Measuring automatic associations: Validation of algorithms for the Implicit Association Test (IAT) in a laboratory setting

J. Behav. Ther. & Exp. Psychiat. 44 (2013) 105–113

Contents lists available at SciVerse ScienceDirect

Journal of Behavior Therapy and Experimental Psychiatry

journal homepage: www.elsevier.com/locate/jbtep

Measuring automatic associations: Validation of algorithms for the Implicit Association Test (IAT) in a laboratory setting

Klaske A. Glashouwer [a,*], Fren T.Y. Smulders [b], Peter J. de Jong [a], Anne Roefs [b], Reinout W.H.J. Wiers [c]

[a] Department of Clinical Psychology, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands
[b] Department of Clinical Psychological Science, Faculty of Psychology and Neuroscience, Maastricht University, The Netherlands
[c] Faculty of Social and Behavioural Sciences – Developmental Psychology, University of Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 30 January 2012
Received in revised form
6 July 2012
Accepted 30 July 2012

Keywords:
IAT
Automatic associations
Algorithm
D-measure
Laboratory setting
Psychopathology

ABSTRACT

Background and objectives: In their paper, "Understanding and using the Implicit Association Test: I. An improved scoring algorithm", Greenwald, Nosek, and Banaji (2003) investigated different ways to calculate the IAT-effect. However, up to now, it remained unclear whether these findings – based on internet data – also generalize to laboratory settings. Therefore, the main goal of the present study was to cross-validate scoring algorithms for the IAT in a laboratory setting, specifically in the domain of psychopathology.
Methods: Four known IAT algorithms and seven alternative IAT algorithms were evaluated on several performance criteria in the large-scale laboratory sample of the Netherlands Study of Depression and Anxiety ($N = 2981$) in which two IATs were included to obtain measurements of automatic self-anxious and automatic self-depressed associations.
Results and conclusions: Results clearly demonstrated that the D_{2SD}-measure and the D_{600}-measure as well as an alternative algorithm based on the correct trials only (D_{noEP}-measure) are suitable to be used in a laboratory setting for IATs with a fixed order of category combinations. It remains important to further replicate these findings, especially in studies that include outcome measures of more spontaneous kinds of behaviors.

1. Introduction

During the past two decades, an increased interest for implicit associations has also spread to the field of psychopathology (e.g., De Houwer, 2002) with the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) as one of the most frequently used measurement instruments. This kind of research is inspired by recent information-processing models that emphasize the importance to distinguish between more explicit and more automatically activated cognitions. Both types of cognitions are believed to have different functional qualities (e.g. Gawronski & Bodenhausen, 2006) and influence different kinds of behaviors. While explicit cognitions are assumed to predict more deliberate, controlled behaviors, implicit associations are thought to predict behaviors when these behaviors are uncontrollable. Sometimes behaviors are inherently uncontrollable. At other times, people don't have the capacity to control their behaviors, do not feel the need to control, or are not aware of the influence of their implicit associations (e.g., Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005). The latter kinds of behaviors are also critically involved in psychopathology where patients often report symptoms being unpredictable and uncontrollable (e.g., Mayer, Merckelbach, & Muris, 2000).

Despite the frequent use of the IAT within psychopathology research (review: Roefs et al., 2011), there are still several unsolved methodological and conceptual issues regarding what the IAT actually measures (e.g., Conrey et al., 2005; Fiedler, Messner, & Bluemke, 2006; Klauer, Schmitz, Teige-Mocigemba, & Voss, 2010) and to what extent IAT-effects really reflect 'implicit' or 'automatic' cognitive processes (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009). The present paper focuses on one specific methodological issue, namely how response latencies of the IAT can be transformed into a meaningful outcome measure, i.e. the so-called "IAT-effect". Our objective is to maintain as much information as possible gathered with the IAT that is meaningful and useful (i.e. reflecting the implicit associations of interest) and at the same time leave out misinformation considered to be 'noise' (for an example of an IAT design see block 1–7 of Table 1).

In their paper, "Understanding and using the Implicit Association Test: I. An improved scoring algorithm", Greenwald, Nosek, and

* Corresponding author. Tel.: +31 50 3636390; fax: +31 50 3637602.
E-mail addresses: k.a.glashouwer@rug.nl (K.A. Glashouwer), f.smulders@maastrichtuniversity.nl (F.T.Y. Smulders), p.j.de.jong@rug.nl (P.J. de Jong), a.roefs@maastrichtuniversity.nl (A. Roefs), R.W.H.J.Wiers@uva.nl (R.W.H.J. Wiers).

0005-7916/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jbtep.2012.07.015

Banaji (2003) thoroughly investigated different ways to calculate the IAT-effect. They showed that in large datasets collected through the internet the so-called D-measures perform best. However, it is still unclear whether these findings also generalize to laboratory settings, which are common in psychopathology research. Therefore, the main goal of the present study was to cross-validate scoring algorithms for the IAT in a laboratory setting in the domain of psychopathology. Given the dominant influence of the IAT in this field, a better understanding of its scoring procedures seems crucial.

In contrast to laboratory settings, internet studies almost completely lack experimental control, which could lower the commitment of participants to the task and by this, create more lapses of attention. Because short periods of inattention probably increase both the average and the variability of reaction times (RTs), it might be that the superior performance of D-measures (that correct for variability by dividing by the inclusive SD) is limited to situations without experimental control. Consistent with the suggestion that the D-measure might in fact be suboptimal for indexing IAT-effects in laboratory studies, some studies in the alcohol-domain (Wiers, Rinck, Kordts, Houben, & Strack, 2010; Wiers, van den Luitgaarden, van den Wildenberg, & Smulders, 2005), found expected changes in IAT-scores as a result of a cognitive behavioral intervention using the original algorithm, but not with D-measures. This could be related to the more controlled lab-circumstances, to the within-subjects designs used (which were not used in the original validation study of Greenwald et al., 2003), or just reflect chance findings, given the relatively small sample sizes of these studies. In any case, currently, most studies using the IAT only report results of D-measures, which makes it hard to compare the performance of various scoring algorithms across different settings.

How should we judge which IAT algorithm performs best? Greenwald et al. (2003) formulated several criteria on which they compared the performance of different IAT algorithms: correlation with explicit measures; correlation with average response latency; internal consistency; sensitivity to undesired influence of order effects of the combined task; resistance to the effect of prior IAT experience; effect size; and magnitude of the implicit–explicit path. They identified the correlation with the explicit equivalent as one of the most important performance criteria based on the assumption that implicit and explicit measures share one underlying attitude. However, this assumption can be questioned on the basis of current dual process models (e.g., Gawronski & Bodenhausen, 2006; Strack & Deutsch, 2004). According to these models, input of the associative network forms the basis of propositional reasoning leading to explicit cognitions. Implicit associations and explicit cognitions will often work synchronously, and in these cases higher correlations of IAT-effects with explicit cognitions would be favorable, since it indicates less random measurement error of the IAT. However, in some cases explicit cognitions may differ from implicit associations and lead to behavioral outcomes that differ from the associative pathway. The latter implies that a high correlation between implicit associations and explicit cognitions it is not by definition favorable. IATs measuring implicit associations that conflict with explicit cognitions will show lower correlations with explicit attitudes. In these situations, selecting for scores that have high correlations will decrease the divergent validity of the IAT. Nevertheless, we included this criterion in the present paper to be able to compare the results with the prior work of Greenwald and colleagues.

As a second important performance criterion, the correlation with general response speed was used, because people with a slower overall response tendency generally show larger IAT-effects. Furthermore, conceptually unrelated IAT-effects show substantial correlations (Back, Schmukle, & Egloff, 2005; Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007; McFarland & Crouch, 2002; Mierke & Klauer, 2003). These findings might be explained by individual differences in general cognitive abilities, since IAT-effects were found to be the result of task-switching abilities and (to a smaller extent) working memory capacity (Klauer et al., 2010; Mierke & Klauer, 2003). Although association strengths might affect the ease of task switching in the IAT, it could also be that general cognitive abilities which are unrelated to the implicit associations of interest have a confounding influence on IAT-effects. In the present study, we did not have separate measures of cognitive abilities. However, in line with Greenwald et al. (2003), we looked at the correlation with general response speed hypothesizing that IAT-effects that show smaller correlations with general response speed seem to be preferable.

In addition to the criteria that Greenwald et al. (2003) formulated, we included predictive validity as an additional criterion by examining the ability of the IAT to predict relevant outcome measures (Blanton & Jaccard, 2006). Greenwald, Poehlman, Uhlmann, and Banaji (2009) already conducted a large meta-analysis on the predictive validity of the IAT. Although this meta-analysis included IAT-scoring method as a methodological moderator, the performance of different IAT algorithms with respect to predictive validity could not be compared within one sample.

To summarize, the present study is an extension of the work of Greenwald et al. (2003) who started examining how response latencies of the IAT can be transformed into a meaningful outcome measure. Our main goal was to cross-validate scoring algorithms for the IAT in a laboratory setting in the domain of psychopathology. Therefore, four known IAT algorithms will be compared on several performance criteria, including predictive validity. In addition, this study will explore the performance of seven alternative IAT algorithms. It is very hard to find laboratory datasets that are sufficiently large to achieve the required power for the purpose of the present enterprise. Fortunately, we had access to the unique, large-scale sample of Netherlands Study of Depression and Anxiety (NESDA; $N = 2981$) in which participants carried out two IATs in a laboratory setting. The IATs were designed to measure implicit self-anxious and implicit self-depressed associations, respectively (cf. Egloff & Schmukle, 2002). Data were collected among both patients and non-clinical controls and the assessment was repeated after two years (Penninx et al., 2008).

2. Method

2.1. IAT algorithms

Eleven different IAT algorithms were tested: four were the same as in the study of Greenwald et al. (2003): D_{2SD}-measure, D_{600}-measure, C_1-measure and C_3-measure. Furthermore, seven alternative algorithms were tested: D_{noEP}-measure, D_{noSD}-measure, $D_{noSD+log}$-measure, GRS-measure, d-measure, S-measure and P-measure. In line with the recommendations of Greenwald and colleagues, we decided to apply two basic principles to all the algorithms: 1) participants with more than 10% of the RTs below 300 milliseconds (ms) were discarded from the analyses; 2) error trials were replaced with mean reaction times of correct responses in the block in which the error occurred plus a penalty of 600 ms. Because the present IAT design did not record the second correct response after a mistake, no built-in error penalty could be used. The only exceptions to the second rule were the D_{2SD}-measure and the D_{noEP}-measure that did not include an error penalty of 600 ms (see description below). In addition, subjects with high error rates or scores diverging more than 4 standard deviations (SDs) from the mean were discarded from the analyses (see 'Missing data and

Table 1
Arrangement of Implicit Association Test blocks.

Block	No. of trials	Function	Labels assigned to left-key response	Labels assigned to right-key response
1	20	Practice	me	other
2	20	Practice	anxious	calm
3	20	Practice	me + anxious	other + calm
4	60	Test	me + anxious	other + calm
5	20	Practice	calm	anxious
6	20	Practice	me + calm	other + anxious
7	60	Test	me + calm	other + anxious
8	20	Practice	depressed	elated
9	20	Practice	me + depressed	other + elated
10	60	Test	me + depressed	other + elated
11	20	Practice	elated	depressed
12	20	Practice	me + elated	other + depressed
13	60	Test	me + elated	other + depressed

construction of groups'). Main characteristics of the measures and rationales behind the new measures will be briefly discussed below. A detailed overview of the characteristics of the algorithms can be found in Tables 2 and 3.

2.1.1. D_{2SD}-measure

In the D_{2SD}-measure, both the practice and the test trials were included as well as the first trials of each of these blocks. RTs above 10,000 ms were excluded before mean RTs were calculated for all blocks. Error trials were replaced with mean reaction times of correct responses in the block in which the error occurred plus a penalty of twice the SD of correct responses in the block in which the error occurred. For practice trials and test trials separately, the difference scores between the congruent and incongruent blocks were divided by the SDs of these blocks, i.e. the difference score of the practice blocks was divided by the SD based on all practice trials and the difference score of the test blocks was divided by the SD based on all test trials. Then, the unweighted mean of both difference scores was calculated.

2.1.2. D_{600}-measure

The D_{600}-measure is similar to the D_{2SD}-measure apart from the error penalty which is 600 ms for the D_{600}-measure. In addition to the D_{2SD}-measure and the D_{600}-measure, Greenwald et al. (2003) calculated four other D-measures. However, in the present study no built-in error penalty could be used, which means that D_1-measure and D_2-measure could not be calculated. Greenwald and colleagues state that "for the four D-measures (D_3, D_4, D_5, D_6) that replaced error latencies with computed penalties, there were virtually no differences between the two measures that deleted latencies below 400 ms (D_5 and D_6) and the two that did not (D_3 and D_4)." Consequently, we decided to use the D_3-measure (that we refer to as 'D_{2SD}-measure') and the D_4-measure (that we refer to as 'D_{600}-measure') to be able to include as many of the trials as possible.

2.1.3. C_1-measure

The C_1-measure ('C' for 'conventional') was only based on the test trials with excluding the first two trials of the blocks as 'warm-ups'. RTs below 300 ms and above 3000 ms were recoded to respectively 300 and 3000 ms. RTs were log-transformed before the mean score for each block was calculated.

2.1.4. C_3-measure

The C_3-measure was almost similar to the C_1-measure. The only difference was that both the test and the practice trials were included in the algorithm, instead of the test trials only. The unweighted mean of both difference scores (practice and test) formed the C_3-measure.

2.1.5. D_{noEP}-measure

To investigate whether adding an error penalty is indeed improving the IAT-effect, D_{noEP}-measure was included based on the correct trials only. The remaining transformations were the same as with the D_{2SD}-measure and the D_{600}-measure.

2.1.6. D_{noSD}-measure

To investigate whether dividing by the SD was indeed improving the IAT-effect, or whether the other ingredients of the D-measure caused this effect, the D-measure was tested without dividing by the SD. The remaining transformations were the same as with the D_{600}-measure.

2.1.7. $D_{noSD+log}$-measure

This measure was very similar to the D_{noSD}-measure and the C_3-measure. The difference with the D_{noSD}-measure is that the RTs were log-transformed before averaging and the difference with the C_3-measure is that RTs above 10,000 ms were excluded instead of recoding RTs below 300 ms and above 3000 ms.

2.1.8. GRS-measure

If correcting for general response speed is an important element of the IAT algorithm, it might be useful to directly divide the difference score by the general response speed (GRS), instead of correcting for it by dividing by the SD. The GRS-measure was otherwise similar to the D_{600}-measure, but instead of dividing each difference score by its SD, the unweighted mean difference scores of the practice and test blocks were divided by general response speed. GRS was defined here as the mean RT of the single target practice trials.

Table 2
Characteristics of known IAT algorithms.

	D_{2SD}-measure (Greenwald et al., 2003)	D_{600}-measure (Greenwald et al., 2003)	C_1-measure (Greenwald et al., 1998)	C_3-measure (Greenwald et al., 2003)
Which trials?	Practice and test Include first trials	Practice and test Include first trials	Test only Exclude first trials	Practice and test Include first trials
Treatment extremes	Exclude trials > 10,000 ms	Exclude trials > 10,000 ms	Recode RTs < 300 ms and >3000 ms	Recode RTs < 300 ms and >3000 ms
Error penalty	2 SD	600 ms	600 ms	600 ms
Latency transformation	None	None	Natural log-transformation	Natural log-transformation
Other transformation	Divide practice and test difference scores by inclusive SD, before taking unweighted mean	Divide practice and test difference scores by inclusive SD, before taking unweighted mean	None	Unweighted mean practice and test effect

Note. In all algorithms, subjects with more than 10% of their responses below 300 ms were excluded from analyses.

Table 3
Characteristics of new IAT algorithms.

	D_{noEP}-measure	D_{noSD}-measure	$D_{noSD+log}$-measure	GRS-measure	d-measure	S-measure	P-measure
Which trials?	Practice and test Include first trials	Practice and test Include first trials	Practice and test Include first trials	Practice and test Include first trials	Practice and test Include first trials	Practice and test Include first trials	Practice only Include first trials
Treatment extremes	Exclude trials > 10,000 ms	Exclude trials > 10,000 ms	Exclude trials > 10,000 ms	Exclude trials > 10,000 ms	Exclude trials > 10,000 ms	Exclude trials > 10,000 ms	Exclude trials > 10,000 ms
Error penalty	Correct trials only	600 ms	600 ms	600 ms	600 ms	600 ms	600 ms
Latency transformation	None	None	Natural log-transformation	None	None	None	None
Other transformations	Divide practice and test difference scores by inclusive SD, before taking unweighted mean	Unweighted mean practice and test effect	Unweighted mean practice and test effect	Unweighted mean practice and test effect before dividing this mean by GRS[a]	Calculate Cohen's d[b] for practice and test blocks, before taking unweighted mean	Unweighted mean practice and test effect, before recoding effects > 0 into +1 and <0 into −1	None

Note. GRS, general response speed; S, sign; P, practice. In all algorithms, subjects with more than 10% of their responses below 300 ms were excluded from analyses.
[a] GRS is defined as the mean RT of the single target practice trials in blocks 1,2, and 5.
[b] Cohen's d is defined as $d = \frac{\overline{x_1} - \overline{x_2}}{s}$, $s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2}}$.

2.1.9. d-measure

Cohen's *d* is a widely used way to calculate effect size. Therefore, it makes sense to test the performance of this measure as well. The d-measure was again similar to the D_{600}-measure, but now Cohen's *d* was calculated. For the d-measure, the difference score was divided by the standard deviation. This was done for the test and practice trials separately, before the unweighted mean was taken. The difference between the present d-measure and the D_{600}-measure is that the standard deviation in the denominator of the D_{600}-measure is computed from all scores in both conditions (congruent and incongruent blocks taken together), ignoring the condition membership of each score. By contrast, the standard deviation used in computing the effect size of Cohen's *d* is a within-condition standard deviation (congruent and incongruent blocks separately).

2.1.10. S-measure

S stands for 'sign'. In this algorithm, practice as well as test trials were included. After the unweighted mean of the difference scores was calculated, these effects were being recoded in a dichotomous variable according to their sign. Values above zero were recoded into +1 and values below zero were recoded into −1. The advantage of this measure could be that it is quite robust against how exactly the IAT-effect was calculated. The S-measure only shows which combination (congruent or incongruent) someone finds more difficult. Therefore, task-specific variance probably will have less influence, because the chance that it will change the sign is small.

2.1.11. P-measure

Greenwald et al. (2003) found that correlations with explicit measures were higher for the IAT measures based on the practice blocks than for the measures based on test blocks. Perhaps this has to do with the tendency of IAT-effects to decrease when individuals have more experience conducting the IAT (e.g., Greenwald & Nosek, 2001; Wiers et al., 2005). It might be that the most important information can be obtained from the first few trials. Following this perspective, we decided to test an algorithm that only includes the practice trials, i.e. the P-measure was calculated as the difference score between the congruent and incongruent practice blocks.

2.2. Criteria and corresponding data analyses

The IAT algorithms were examined according to the following six performance criteria: 1) IAT correlations with explicit measures (high values desired); 2) Correlations of IAT with response latency (approaching zero desired); 3) Internal consistency (high values desired); 4) Resistance to undesired influence of prior IAT experience; 5) IAT-effect size (high values desired); 6) Predictive validity (high values desired).

2.2.1. IAT correlations with explicit measures

Correlations between IAT measures and explicit equivalent measures were calculated.

2.2.2. Correlations of IAT with response latency

We used the mean response speed of the four combined blocks as measure of general response speed (GRS). Greenwald et al. (2003) calculated GRS without excluding trials above 10,000 ms. However, we did not want extreme RTs to influence GRS and excluded trials above 10,000 ms before calculating GRS. Correlations were calculated between the absolute IAT-effects and GRS.

2.2.3. Internal consistency

Greenwald et al. (2003) defined the internal consistency of the IAT as the correlation between IAT-effects based on two mutually exclusive subsets. As subsets they used the practice- and test trials, and when this was not possible, the first and the second half of the blocks were used. A higher correlation points to a stronger internal consistency and indicates a better measure. However, during the administration of the IAT, usually learning effects occur, and therefore, the correlation between IAT-effects based on test and practice trials (or first and second block halves) could in fact underestimate the internal consistency. Therefore, we applied a slightly different definition and calculated the Spearman–Brown corrected correlations between IAT-effects based on two mutually exclusive subsets of 'odd and even trials' (test-halves were based on trials 1, 2, 5, 6, 9, 10 etc. vs. 3, 4, 7, 8, 11, 12 etc.).

2.2.4. Resistance to undesired influence of prior IAT experience

Additionally, Greenwald et al. (2003) looked at the correlation of the IAT-effect with prior IAT experience. IAT-effects tend to decrease with the number of IATs presented to a participant; therefore the algorithm should reduce this influence as much as possible. We tested this in the healthy control group, with repeated measures ANOVA in which Time was included as within-subject factor. The effect of time (η^2) is preferably small.

2.2.5. IAT-effect size

The IAT should be sensitive enough to detect individual or group differences. This means that an algorithm that maximizes IAT-effect sizes is preferable. Sensitivity for differences between groups (anxious/depressed vs. controls) was evaluated using *t*-tests and one-sample *t*-tests were used to test which algorithm was most sensitive to pick up differences between the congruent and incongruent condition of the IAT. For all *t*-tests Cohen's *d* was calculated.

2.2.6. Predictive validity

To test the predictive validity, multiple regression analyses were conducted for the different algorithms separately using depressive and anxiety symptoms as dependent variables. In the Result section R^2 is reported. A description of the symptom measures can be found below.

2.3. Study sample

The present study was carried out in the context of the Netherlands Study of Depression and Anxiety (NESDA, Penninx et al., 2008), a multi-center, ongoing cohort study, designed to examine the long-term course and consequences of anxiety and depressive disorders. A total of 2981 persons aged 18 through 65 were included, including healthy controls, individuals at risk because of prior episodes, sub threshold symptoms or family history, and individuals with a current first or recurrent depressive and/or anxiety disorder. The inclusion was restricted to Major Depressive Disorder, Dysthymia, General Anxiety Disorder, Panic Disorder, Social Phobia, and Agoraphobia, because these disorders are relatively homogeneous in phenotype and are found across different health care settings. Recruitment of respondents took place in the general population, in general practices, and in mental health care institutions. General exclusion criteria were having a primary clinical diagnosis of a psychiatric disorder not subject of NESDA which would importantly affect course trajectory (i.e., psychotic disorder or bipolar disorder) and not being fluent in Dutch. The present study concerns baseline and 2-year follow-up measurements conducted from September 2004 until April 2009. The study protocol was approved centrally by the Ethical Review Board of VU Medical Center Amsterdam and subsequently by local review boards of each participating center/institute, and all participants provided written informed consent.

After two years, a face-to-face follow-up assessment was conducted with a response of 87.1% ($N = 2596$). The attrition rate was relatively low (12.9%) as compared with other psychiatric epidemiological studies (e.g., Eich et al., 2003; de Graaf, Bijl, Smit, Ravelli, & Vollebergh, 2000). Non-response was significantly higher among those with younger age, lower education, non-European ancestry, and depressive disorder, but was not associated with gender or anxiety disorder (Lamers et al., 2012). The presence of depressive or anxiety disorders was established with the Composite International Diagnostic Interview (CIDI; WHO version 2.1) which classifies diagnoses according to DSM-IV criteria (American Psychiatric Association, 2000).

2.4. Measures

2.4.1. Implicit Association Tests

The Implicit Association Test (IAT) is a computerized reaction time task originally designed by Greenwald et al. (1998) to measure the relative strengths of automatic associations between two contrasted target concepts and two attribute concepts. Words from all four concept categories appear in the middle of a computer screen and participants are instructed to sort them with a left or right response key. The premise here is that the sorting becomes easier when a target and attribute that share the same response key are strongly associated than when they are weakly associated. IAT stimuli are presented randomly, but switch from target categories to attribute categories on every trial. The category labels are visible in upper left and right-hand corners of the screen during the whole task. For both IATs target labels were 'me' and 'others'. Following the design of Egloff and Schmukle (2002), an anxiety IAT was constructed with attribute labels 'anxious' and 'calm'. Analogously, attribute labels were 'depressed' and 'elated' for the depression IAT. Each category consisted of five stimuli (see Appendix A). Attribute stimuli of the anxiety IAT were the same self-descriptors as used by Egloff and Schmukle (2002) who based their IAT on trait anxiety. Furthermore, we designed a self-depressed IAT in an equivalent way and selected trait self-descriptors of depressed persons that were also used in previous work on attentional bias in (remitted) depression (e.g., McCabe, Gotlib, & Martin, 2000). Both IATs consisted of two critical test blocks that were preceded by practice blocks (see Table 1). The order of category combinations was fixed across participants to reduce method variance.

2.4.2. Explicit self-associations

To obtain explicit self-associations equivalently to the implicit self-associations, participants rated all IAT attribute stimuli on a 5-point scale (1 = hardly/not at all, 5 = very much). The instruction was "For each word please indicate to what extent you think it generally applies to you" (cf. Back, Schmukle, & Egloff, 2009).

2.4.3. Questionnaire data

Severity of anxiety symptoms was measured with the 21-item Beck Anxiety Inventory (BAI; Beck, Epstein, Brown, & Steer, 1988), whereas fearful avoidance behavior was measured using the 15-item Fear Questionnaire (FQ; Marks & Mathews, 1979). Severity of depressive symptoms was measured with the 30-item Inventory of Depressive Symptoms self-report version (IDS-SR; Rush, Gullion, Basco, & Jarrett, 1996). Total scale scores were used for all questionnaires.

2.5. Procedure

Baseline and follow-up assessments were similar and lasted between 3 and 5 h. During assessments, other measurements were collected as well, but these are not of interest for the present study (for a detailed description, see Penninx et al., 2008). Each participant completed the anxiety IAT, followed by the depression IAT. After that, participants deliberately rated attribute words that were used in the IATs. Respondents were compensated with a €15, gift certificate and travel expenses.

2.6. Missing data and construction of groups

In addition to the 'regular' attrition of the NESDA study, there was extra attrition for the IAT. Sometimes individuals were willing to participate in the follow-up assessment, but were measured at home or via the telephone, resulting in a loss of IAT data. In total, IAT data and explicit self-associations for 129 participants were missing at *t*1 and 564 were missing at *t*2. Participants with more than 10% of the trials below 300 ms (IAT anxiety: $n_{t1} = 7$, $n_{t2} = 1$; IAT depression: $n_{t1} = 7$, $n_{t2} = 1$) and with high error rates (>33.3%; IAT anxiety: $n_{t1} = 16$, $n_{t2} = 8$; IAT depression: $n_{t1} = 19$, $n_{t2} = 6$) were discarded from all analyses. Consequently, at *t*1, the sample consisted of 2829 participants for IAT anxiety and 2826 participants for IAT depression. At t2, the sample consisted of 2023 participants for IAT anxiety and 2025 participants for IAT depression. Furthermore, for the analyses of criterion 6 'predictive validity' 35 individuals

Table 4
Performance of 11 IAT algorithms on 6 criteria in the NESDA sample, IAT anxiety.

Performance criteria	Known algorithms				New algorithms						
	D_{2SD}	D_{600}	C_1	C_3	D_{noEP}	D_{noSD}	$D_{noSD+log}$	GRS	d	S	P
1. Correlation w explicit equivalent											
T1 EA anxiety	.376	.373	.325	.354	.370	.298	.349	.335	.363	.306	.280
T2 EA anxiety	.358	.357	.297	.336	.358	.288	.335	.329	.337	.264	.215
2. Correlation w response speed											
T1	.094	.091	.208	.207	.048	.509	.286	.269	.134		.504
T2	.034	.042	.213	.192	−.003	.484	.234	.240	.061		.454
3. Internal consistency											
T1	.914	.919	.909	.938	.908	.897	.934	.896	.864	.728	.821
T2	.906	.910	.907	.927	.894	.879	.923	.881	.856	.717	.777
4. Prior IAT experience (in controls)	.026	.022	.002	.006	.023	.000	.004	.008	.017	.010	.000
5. IAT-effect size											
T1 One-sample *t*-test	.572	.590	.721	.567	.572	.450	.551	.506	.345	.518	.322
T1 Group differences	.805	.798	.648	.736	.791	.593	.724	.713	.766	.659	.561
T2 One-sample *t*-test	.835	.849	.941	.825	.820	.700	.809	.738	.620	.756	.530
T2 Group differences	.899	.893	.727	.840	.903	.712	.841	.821	.829	.661	.690
6. Predictive validity (R^2)											
T1 BAI	.114	.110	.088	.102	.106	.074	.098	.092	.103	.080	.063
T1 FQ	.082	.081	.058	.071	.078	.049	.068	.063	.074	.054	.046
T2 BAI	.091	.089	.058	.076	.088	.055	.074	.074	.080	.058	.053
T2 FQ	.080	.078	.050	.067	.076	.046	.067	.067	.066	.049	.049

Note. Abbreviations for 11 measures and 6 performance criteria are explained in detail in the Method section. IAT = Implicit Association Test; BAI = Beck Anxiety Inventory; FQ = Fear Questionnaire; EA = Explicit Association.

were discarded because of missing data on the BAI and FQ (100 missings at *t*2) and 39 on the IDS-SR (93 missings at *t*2). Criterion 4 'Resistance to undesired influence of prior IAT experience' was tested in the group of control participants that did not have a disorder during or in between baseline and follow-up ($n = 817$). Finally, for criterion 5 'IAT-effect size' the sensitivity for differences between groups was tested (anxious group: $n_{t1} = 507$, $n_{t2} = 438$; depressed group: $n_{t1} = 280$, $n_{t2} = 272$; controls: $n_{t1} = 643$, $n_{t2} = 437$). The control group was smaller in the latter analyses, because we additionally excluded individuals that had a prior depressive or anxiety disorder (cf. Glashouwer & de Jong, 2010).

In addition, subjects with scores diverging more than 4 SDs from the mean were discarded from the analyses (IAT anxiety: D_{2SD}-measure: $n_{t1} = 0$, $n_{t2} = 0$; D_{600}-measure: $n_{t1} = 0$, $n_{t2} = 0$; C_1-measure: $n_{t1} = 10$, $n_{t2} = 3$; C_3-measure: $n_{t1} = 1$, $n_{t2} = 3$; D_{noEP}-measure: $n_{t1} = 0$, $n_{t2} = 0$; $D_{no\ SD}$-measure: $n_{t1} = 20$, $n_{t2} = 15$; $D_{no\ SD+log}$-measure: $n_{t1} = 3$, $n_{t2} = 4$; GRS-measure: $n_{t1} = 8$, $n_{t2} = 5$; D-measure: $n_{t1} = 21$, $n_{t2} = 14$; S-measure: $n_{t1} = 0$, $n_{t2} = 0$; P-measure:

Table 5
Performance of 11 IAT algorithms on 6 criteria in the NESDA sample, IAT depression.

Performance criteria	Known algorithms				New algorithms						
	D_{2SD}	D_{600}	C_1	C_3	D_{noEP}	D_{noSD}	$D_{noSD+log}$	GRS	d	S	P
1. Correlation w explicit equivalent											
T1 EA depression	.383	.379	.314	.371	.384	.324	.364	.361	.373	.315	.305
T2 EA depression	.356	.358	.291	.337	.366	.299	.337	.335	.329	.279	.244
2. Correlation w response speed											
T1	.080	.078	.205	.190	.030	.467	.257	.251	.110		.499
T2	.057	.069	.235	.234	.029	.455	.262	.224	.078		.444
3. Internal consistency											
T1	.861	.865	.855	.889	.847	.829	.885	.840	.801	.660	.725
T2	.845	.845	.866	.886	.832	.835	.881	.839	.765	.646	.696
4. Prior IAT experience (in controls)	.012	.011	.002	.003	.010	.000	.002	.008	.011	.009	.000
5. IAT-effect size											
T1 One-sample *t*-test	.574	.584	.487	.587	.595	.493	.579	.533	.393	.529	.443
T1 Group differences	.825	.822	.623	.758	.855	.674	.743	.708	.830	.676	.621
T2 One-sample *t*-test	.759	.768	.588	.744	.761	.639	.742	.689	.628	.680	.634
T2 Group differences	.966	.960	.793	.906	.997	.748	.894	.870	.890	.752	.657
6. Predictive validity (R^2)											
T1 IDS	.118	.116	.079	.108	.120	.080	.104	.102	.107	.082	.069
T2 IDS	.087	.088	.064	.076	.093	.053	.073	.074	.076	.060	.041

Note. Abbreviations for 11 measures and 6 performance criteria are explained in detail in the Method section. IAT = Implicit Association Test; IDS = Inventory of Depressive Symptomatology; EA = Explicit Association.

Table 6
Average performance of 11 IAT algorithms on 6 criteria in the NESDA sample.

Performance criteria	Known algorithms				New algorithms						
	D_{2SD}	D_{600}	C_1	C_3	D_{noEP}	D_{noSD}	$D_{noSD+log}$	GRS	d	S	P
1. Correlation w explicit equivalent	.368	.367	.307	.350	**.370**	.300	.346	.340	.351	.291	.261
2. Correlation w response speed	.066	.070	.215	.206	**.026**	.479	.260	.246	.096		.475
3. Internal consistency	.882	.885	.885	**.910**	.870	.860	.906	.864	.822	.688	.755
4. Prior IAT experience (in controls)	.019	.017	.002	.005	.017	**.000**	.003	.008	.014	.010	**.000**
5. IAT-effect size											
One-sample *t*-test	.685	**.698**	.684	.681	.687	.571	.670	.617.	.497	.621	.482
Group differences	.874	.868	.698	.810	**.887**	.682	.801	.778	.829	.687	.632
6. Predictive validity (R^2)	**.095**	.094	.066	.082	.094	.060	.081	.079	.084	.064	.054

Note. Abbreviations for 11 measures and 6 performance criteria are explained in detail in the Method section. IAT = Implicit Association Test. **Bold** = best mean performance. Underlined: second and third best performance.

$n_{t1} = 15$, $n_{t2} = 12$; IAT depression: D_{2SD}-measure: $n_{t1} = 0$, $n_{t2} = 3$; D_{600}-measure: $n_{t1} = 2$, $n_{t2} = 3$; C_1-measure: $n_{t1} = 8$, $n_{t2} = 3$; C_3-measure: $n_{t1} = 3$, $n_{t2} = 4$; D_{noEP}-measure: $n_{t1} = 1$, $n_{t2} = 1$; $D_{no\ SD}$-measure: $n_{t1} = 18$, $n_{t2} = 11$; $D_{no\ SD+log}$-measure: $n_{t1} = 5$, $n_{t2} = 4$; GRS-measure: $n_{t1} = 7$, $n_{t2} = 7$; d-measure: $n_{t1} = 13$, $n_{t2} = 9$; S-measure: $n_{t1} = 0$, $n_{t2} = 0$; P-measure: $n_{t1} = 9$, $n_{t2} = 8$).

3. Results

The results of both IATs separately can be found in Tables 4 and 5. In Table 6 the mean performance of the different algorithms are shown.

3.1. IAT correlations with explicit measures

Overall, the D_{noEP}-measure showed the highest correlation with explicit equivalents. The correlations of the D_{2SD}-measure and the D_{600}-measure were closest to that of the D_{noEP}-measure.

3.2. Correlations of IAT with response latency

The D_{noEP}-measure showed consistently the lowest mean correlation with GRS. The performance of the D_{2SD}-measure and the D_{600}-measure were closest to that of the D_{noEP}-measure.

3.3. Internal consistency

The C_3-measure consistently showed the highest internal consistency. The internal consistencies of the D_{600}-measure, the C_1-measure and the $D_{noSD+log}$-measure were closest to that of the C_3-measure.

3.4. Resistance to undesired influence of prior IAT experience

The D_{noSD}-measure and the P-measure were consistently the most resistant to undesired influence of prior IAT experience. The C_1-measure and the $D_{noSD+log}$-measure performed second and third best on this criterion.

3.5. IAT-effect size

3.5.1. One-sample t-test

The D_{600}-measure showed the greatest mean effect size on the one-sample *t*-tests. The effect sizes of the D_{noEP}-measure and the D_{2SD}-measure were closest to that of the D_{600}-measure.

3.5.2. Group differences

The D_{noEP}-measure showed the greatest mean effect size on the between sample *t*-tests. The effect sizes of the D_{2SD}-measure and the D_{600}-measure were closest to that of the D_{noEP}-measure.

3.6. Predictive validity

The D_{2SD}-measure consistently showed the highest predictive validity. The predictive validities of the D_{600}-measure and the D_{noEP}-measure were very close to that of the D_{2SD}-measure.

4. Discussion

The main purpose of the present study was to extend the findings of Greenwald et al. (2003) and validate scoring algorithms for the IAT in a laboratory setting in the domain of psychopathology. Therefore, four known IAT algorithms (D_{2SD}-measure, D_{600}-measure, C_1-measure and C_3-measure; Greenwald et al., 2003) were evaluated on six performance criteria in the large-scale laboratory sample of the NESDA. In line with the study of Greenwald and colleagues, results demonstrated that the D-measures (D_{2SD}-measure and D_{600}-measure) showed higher correlations with explicit equivalents and lower correlations with general response speed than the conventional measures (C_1-measure and C_3-measure), the two criteria that Greenwald and colleagues identified as most important. In addition, the D-measures showed similar internal consistencies as the conventional measures and better performances in terms of effect sizes. In contrast to the study of Greenwald and colleagues, the D-measures seemed somewhat more sensitive to prior IAT experience than the conventional measures, since the effect of Time on IAT-effects was larger for the D-measures than for the conventional measures. However, the effect of Time was still rather small (η^2's = .017 and .019). Finally, the D_{2SD}-measure and the D_{600}-measure showed the best performance for predictive validity, which we added to the original criteria of Greenwald and colleagues.

Since there seem to be considerable differences between data collected via the internet and in the laboratory (e.g., with respect to experimental control or commitment of participants), we hypothesized that scoring algorithms for the IAT might perform differently in both settings and/or in within-subjects designs. However, present findings disprove these hypotheses by replicating prior findings of Greenwald et al. (2003). In line with this study, the results suggest that D-measures showed generally the best performance on the criteria that were used for evaluating the various algorithms. This was not only the case for criteria that were

identified by Greenwald et al. (2003) as most important, i.e. correlation with explicit equivalent and correlation with general response speed, but also for predictive validity which was included as an additional criterion. All in all, the present findings lead to the conclusion that the D_{2SD}-measure and the D_{600}-measure are suitable for use in a laboratory setting in the domain of psychopathology, when using an IAT in which the order of category combination is fixed.

In addition, this study explored the performance of seven alternative IAT algorithms. Although most of the alternative algorithms performed actually (much) worse than the D-measures, the D_{noEP}-measure showed similar or even slightly better performances than the D_{2SD}-measure and the D_{600}-measure. The latter finding suggests that the inclusion of error trials in the IAT algorithm does not seem crucial. It could be that error trials are actually different from accurate trials, but that this difference is not always related to responding too fast (but instead for example by a lapse of attention). In these cases, adding error penalties might actually result in adding noise to the measure. Together, these outcomes point to the conclusion that the success of the D-measures probably stems from the combination of both the division by the inclusive standard deviation and the inclusion of practice trials. By this specific combination of ingredients, up to now, D-measures (D_{2SD}-measure, D_{600}-measure, D_{noEP}-measure) seem to be filtering out the most meaningful information, at least for this specific IAT design. This does not necessarily imply that all other algorithms should be discarded. Future studies will have to illuminate whether the present positive results for the D-measures also hold for laboratory studies with a different design. Therefore, it would be important for coming studies to report results of alternative algorithms next to the D-measure.

4.1. Limitations and considerations

As already mentioned in the introduction, the use of the first criterion, correlation with explicit equivalent can be questioned. This stresses the necessity to look at the performance on the other criteria as well, most importantly predictive validity. Ideally, we would have included outcome measures in our design that are typically assumed to be influenced by implicit associations, e.g., non-verbal behaviors in a stress-task (e.g., Egloff & Schmukle, 2002). Unfortunately, in the present large-scale study there was no room for such labor intensive measures. Consequently, we decided to use questionnaires to measure depressive and anxiety symptoms. Although completing a questionnaire is probably for the larger part not a spontaneous process, it could still capture experiences of behaviors/feelings that occurred spontaneously, which might also explain why we found correlations between implicit associations and these outcome measures. However, by using a self-report measure as outcome measure to test predictive validity, we run the risk of letting in the influence of explicit 'strategic processes' (e.g., Rothermund, Wentura, & De Houwer, 2005; Wentura & Rothermund, 2007). Therefore, an important future research step would be to further validate IAT-scoring algorithms against outcome measures of more spontaneous behaviors, preferably behaviors that are known to be driven primarily by automatic, but not by controlled processes.

Furthermore, because the NESDA sample was not specifically designed for the purpose of the present study, some factors might have influenced the results. First of all, the blocks and the order of both IATs were not counterbalanced between participants. This was done to reduce method variance in consideration of the prospective design of the NESDA. Although other studies chose similar designs (e.g., Asendorpf, Banse, & Mücke, 2002; Schnabel, Banse, & Asendorpf, 2006; Steffens & König, 2006), it hampers the generalizability of the present findings to laboratory studies without fixed blocks/orders of IAT. In addition, the IAT design did not contain a built-in error penalty, which made it impossible to calculate the D_1-measure and the D_2-measure. Greenwald et al. (2003) showed that IAT algorithms with built-in error penalties – if anything – performed slightly better than the other D-measures. Given the similarity in performance of the D_1-measure and D_2-measure to the D_{2SD}-measure and D_{600}-measure that was demonstrated by Greenwald and colleagues, we assume that the D_1-measure and D_2-measure can be used in laboratory settings as well. However, this needs empirical testing, especially since the D_{noEP}-measure, based on correct trials only, showed such good performance.

As a more general issue, we note that the way RTs are used in the context of the IAT differs in an important way from their use in the tradition of Donders (1868), Sternberg (1969), and many others, where they are used to develop a model of the structure of information processing. In this tradition, RTs provide a measure of time duration as a physical property of a mental process, and are usually interpreted on a measurement scale at 'ratio/interval' level. In most of the IAT-literature, however, the aim is not to measure the duration of a process, but rather the strength of an implicit association and this is done indirectly, through its effect on time duration. Therefore, it is unclear whether implicit associations can be measured on an interval scale, or whether the scale should be considered 'ordinal', as often is the case in psychometrics. From this perspective, we should be cautious with interpreting parametric statistics on IAT-effects and future research in this area should focus more on how IAT-effects exactly relate to association strength.

4.2. Conclusion

To summarize, the present study clearly and convincingly demonstrated that the D_{2SD}-measure and the D_{600}-measure as well as an alternative algorithm based on the correct trials only (D_{noEP}-measure) are suitable to be used in a laboratory setting in the domain of psychopathology for IATs with a fixed order of category combinations. However, these findings should be further replicated, especially in studies that include outcome measures of more spontaneous kinds of behaviors. In future studies that make use of the IAT, it would be interesting not only to report results of the D-measure, but also the results of alternative IAT algorithms. Hopefully this will give us even more insight into the optimum use of the Implicit Association Test as a measure for automatic associations.

Declaration of interest

None.

Acknowledgments

The infrastructure for the NESDA study (www.nesda.nl) is funded through the Geestkracht program of the Netherlands Organisation for Health Research and Development (Zon-Mw, grant number 10-000-1002) and is supported by participating universities and mental health care organizations (VU University Medical Center, GGZ inGeest, Arkin, Leiden University Medical Center, GGZ Rivierduinen, University Medical Center Groningen, University of Groningen, Lentis, GGZ Friesland, GGZ Drenthe, Scientific Institute for Quality of Healthcare (IQ Healthcare), Netherlands Institute for Health Services Research (NIVEL) and Netherlands Institute of Mental Health and Addiction (Trimbos)). We thank Bert Hoekzema for technical support; PSY-opleidingen Noord Oost for financial support of the first author.

Appendix A

IAT stimulus words

Me: I, myself, self, my, own
(ik, mezelf, zelf, mijn, eigen)
Others: other, you, they, them, themselves
(ander, jullie, zij, hun, zijzelf)
Anxious: anxious, afraid, nervous, insecure, worried
(angstig, bang, nerveus, onzeker, ongerust)
Calm: calm, balanced, placid, secure, relaxed
(kalm, evenwichtig, rustig, zeker, ontspannen)
Depressed: useless, pessimistic, inadequate, negative, meaningless
(nutteloos, pessimistisch, ongeschikt, negatief, zinloos)
Elated: positive, optimistic, active, valuable, cheerful
(positief, optimistisch, actief, waardevol, opgewekt)

Note. Words are translated from Dutch.

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders*. text revision ed. (4th ed.). Washington, DC: Author

Asendorpf, J. B., Banse, R., & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: the case of shy behavior. *Journal of Personality and Social Psychology, 83*, 380–393.

Back, M. D., Schmukle, S. C., & Egloff, B. (2005). Measuring task-switching ability in the implicit association test. *Experimental Psychology, 52*, 167–179.

Back, M. D., Schmukle, S. C., & Egloff, B. (2009). Predicting actual behavior from the explicit and implicit self-concept of personality. *Journal of Personality and Social Psychology, 97*, 533–548.

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: psychometric properties. *Journal of Consulting and Clinical Psychology, 56*, 893–897.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61*, 27–41.

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: the quad model of implicit task performance. *Journal of Personality and Social Psychology, 89*, 469–487.

De Houwer, J. (2002). The implicit association test as a tool for studying dysfunctional associations in psychopathology: strengths and limitations. *Journal of Behavior Therapy and Experimental Psychiatry, 33*, 115–133.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: a normative analysis and review. *Psychological Bulletin, 135*, 347–368.

Donders, F. C. (1868). Over de snelheid van psychische processen (On the speed of mental processes). *Onderzoekingen gedaan in het Physiologisch Laboratorium der Utrechtsche Hoogeschool, 1868–1869, Tweede reeks, II, 92–120* (W. G. Koster (1969), Trans.). In W. G. Koster, *Attention and performance II. Acta Psychologica, 30*, 412–431.

Egloff, B., & Schmukle, S. C. (2002). Predictive validity of an implicit association test for assessing anxiety. *Journal of Personality and Social Psychology, 83*, 1441–1455.

Eich, D., Ajdacic-Gross, V., Condrau, M., Huber, H., Gamma, A., Angst, J., et al. (2003). The Zurich study: participation patterns and symptom checklist 90-R scores in six interviews, 1979–1999. *Acta Psychiatrica Scandinavica, 108*, 11–14.

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I, " the "A, " and the "T": a logical and psychometric critique of the implicit association test (IAT). *European Review of Social Psychology, 17*, 74–147.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*, 692–731.

Glashouwer, K. A., & de Jong, P. J. (2010). Disorder-specific automatic self-associations in depression and anxiety: results of the Netherlands study of depression and anxiety. *Psychological Medicine, 40*, 1101–1111.

de Graaf, R., Bijl, R. V., Smit, F., Ravelli, A., & Vollebergh, W. A. (2000). Psychiatric and sociodemographic predictors of attrition in a longitudinal study: the Netherlands mental health survey and incidence study (NEMESIS). *American Journal of Epidemiology, 152*, 1039–1047.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology, 74*, 1464–1480.

Greenwald, A. G., & Nosek, B. A. (2001). Health of the implicit association test at age 3. *Zeitschrift für Experimentelle Psychologie, 48*, 85–93.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*, 197–216.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & Banaji, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*, 17–41.

Klauer, K. C., Schmitz, F., Teige-Mocigemba, S., & Voss, A. (2010). Understanding the role of executive control in the implicit association test: why flexible people have small IAT effects. *The Quarterly Journal of Experimental Psychology, 63*, 595–619.

Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Processing components of the implicit association test: a diffusion-model analysis. *Journal of Personality and Social Psychology, 93*, 353–368.

Lamers, F., Hoogendoorn, A., Smit, J. H., van Dyck, R., Zitman, F. G., Nolen, W. A., et al. (2012). Socio-demographic and psychiatric determinants of attrition in the Netherlands study of depression and anxiety (NESDA). *Comprehensive Psychiatry, 53*, 63–70.

Marks, I. M., & Mathews, A. M. (1979). Brief standard self-rating for phobic patients. *Behaviour Research and Therapy, 17*, 263–267.

Mayer, B., Merckelbach, H., & Muris, P. (2000). Self-reported automaticity and irrationality in spider phobia. *Psychological Reports, 87*, 395–405.

McCabe, S. B., Gotlib, I. H., & Martin, R. A. (2000). Cognitive vulnerability for depression: deployment of attention as a function of history of depression and current mood state. *Cognitive Therapy and Research, 24*, 427–444.

McFarland, S. G., & Crouch, Z. (2002). A cognitive skill confound on the implicit association test. *Social Cognition, 20*, 483–510.

Mierke, J., & Klauer, K. C. (2003). Method-specific variance in the implicit association test. *Journal of Personality and Social Psychology, 85*, 1180–1192.

Penninx, B. W. J. H., Beekman, A. T. F., Johannes, H. S., Zitman, F. G., Nolen, W. A., Spinhoven, P., et al. (2008). The Netherlands study of depression and anxiety (NESDA): rationale, objectives and methods. *International Journal of Methods in Psychiatric Research, 17*, 121–140.

Roefs, A., Huijding, J., Smulders, F. T. Y., MacLeod, C. M., de Jong, P. J., Wiers, R. W., et al. (2011). Implicit measures of association in psychopathology research. *Psychological Bulletin, 137*, 149–193.

Rothermund, K., Wentura, D., & De Houwer, J. (2005). Validity of the salience asymmetry account of the implicit association test: reply to Greenwald, Nosek, Banaji, and Klauer (2005). *Journal of Experimental Psychology: General, 134*, 426–430.

Rush, A. J., Gullion, C. M., Basco, M. R., & Jarrett, R. B. (1996). The inventory of depressive symptomatology (IDS): psychometric properties. *Psychological Medicine, 26*, 477–486.

Schnabel, K., Banse, R., & Asendorpf, J. B. (2006). Assessment of implicit personality self-concept using the implicit association test (IAT): concurrent assessment of anxiousness and angriness. *British Journal of Social Psychology, 45*, 373–396.

Steffens, M. C., & König, S. S. (2006). Predicting spontaneous big five behavior with implicit association tests. *European Journal of Psychological Assessment, 22*, 13–20.

Sternberg, S. (1969). The discovery of processing stages: extensions of Donders' method. *Acta Psychologica, 30*, 276–315.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*, 220–247.

Wentura, D., & Rothermund, K. (2007). Paradigms we live by: a plea for more basic research on the implicit association test. In B. Wittenbrink, N. Schwarz, B. Wittenbrink, & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 195–215). New York, NY, US: Guilford Press.

Wiers, R. W., van den Luitgaarden, J., van den Wildenberg, E., & Smulders, F. (2005). Challenging implicit and explicit alcohol-related cognitions in young heavy drinkers. *Addiction, 100*, 806–819.

Wiers, R. W., Rinck, M., Kordts, R., Houben, K., & Strack, F. (2010). Re-training automatic action-tendencies to approach alcohol in hazardous drinkers. *Addiction, 105*, 279–287.